DOCUMENT RESUME

ED 287 868 TM 870 588

AUTHOR Thompson, Bruce
TITLE The Use (and Misuse) of Statistical Significance Testing: Some Recommendations for Improved Editorial Policy and Practice.
PUB DATE Apr 87
NOTE 45p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, DC, April 20-24, 1987).
PUB TYPE Speeches/Conference Papers (150) -- Information Analyses (070) -- Viewpoints (120)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Analysis of Variance; Data Interpretation; Editors; *Effect Size; Error of Measurement; *Hypothesis Testing; Literature Reviews; Reliability; Research Design; *Research Problems; Sample Size; *Scholarly Journals; Statistical Bias; *Statistical Inference; *Statistical Significance
IDENTIFIERS *Editorial Policy; Null Hypothesis

ABSTRACT

This paper evaluates the logic underlying various criticisms of statistical significance testing and makes specific recommendations for scientific and editorial practice that might better increase the knowledge base. Reliance on the traditional hypothesis testing model has led to a major bias against nonsignificant results and to misinterpretation of significant results. A finding of statistical significance does not mean that the null hypothesis is false, since there are many factors affecting statistical significance such as sample size and the measurement reliability of the data. Furthermore, statistical significance alone does not permit evaluation of the importance of a finding. An effect size statistic, such as eta-squared, is more appropriate for this purpose, and editors of scholarly publications should encourage routine reporting of effect sizes. Greater reporting of nonsignificant results should also be encouraged, accompanied by power analyses to estimate Type II error. Nonsignificant results can be meaningful if the study's power to detect an effect was high. Finally, the paper emphasizes the crucial role of replication in separating true effects from Type I errors. The paper integrates analyses and criticisms of statistical practice from a variety of sources--77 references are included. (LPG)

The Use (and Misuse) of Statistical Significance Testing: Some Recommendations for Improved Editorial Policy and Practice

Bruce Thompson

University of New Orleans 70148

and

Louisiana State University Medical Center

Paper presented at the annual meeting of the American Educational Research Association, Washington, DC, April 24, 1987.

There are four chief obstacles to grasping truth, which hinder every man, however learned, and scarcely allow anyone to win a clear title to knowledge; namely, submission to faulty and unworthy authority, influence of custom, popular prejudice, and concealment of our own ignorance accompanied by the ostentatious display of our knowledge.

--Roger Bacon, cited in Andreski, Social sciences as sorcery, 1972, p. 8

Chris: [Sitting back again.] Well, that's part of what's so great about my study. Students in my treatment group had a higher mean score on the individually administered test, and the difference was statistically significant.

Jean: Terrific! Was the difference large enough to be important?

Chris: [Looking puzzled.] Well, as I said, it was statistically significant. You know, that means it wasn't likely to be just a chance occurence. I set the level of significance at 0.05, as my advisor suggested. So a difference that large would occur by chance less than five times in a hundred if the groups weren't really different. An unlikely occurence like that surely must be important.

Jean: Wait a minute, Chris. Remember the other day when you went into the office to call home? Just as you completed dialing the number, your little boy picked up the phone to call someone. So you were connected and talking to one another without the phone ever ringing... What do you suppose the probability would be of his picking up the phone just as you completed dialing?

Chris: Gee, I couldn't even estimate, but it would have to be minuscule.

Jean: Well, that must have been a truly important occurence then?

Chris: . . . . .

--Shaver, Phi Delta Kappan, 1985, p. 58

ABSTRACT

The paper evaluates the logic underlying various criticisms of statistical significance testing and makes specific recommendations for scientific and editorial practice that might better increase the knowledge base. The effects of contemporary significance testing practice on the literature are evaluated. The paper explores why unconscious preferences for certain practices have emerged and why such practices are so impervious to change. The paper attempts to facilitate escape from some of the methodological paradigms that tend to unconsciously govern thinking regarding the processes of scientific inquiry.

Few methodological offerings have sparked more controversy than Sir Ronald Fisher's (1925; 1926) contribution to the logic of null hypothesis testing. The last 30 years have involved periodic efforts (cf. Carver, 1978; Morrison & Henkel, 1970; Selvin, 1957;) by various researchers "to exorcise the null hypothesis" (Cronbach, 1975, p. 124). For example, Shaver (1979, pp. 5-6) has argued that

> The emphasis on statistics and the "test of significance" procedure has resulted in a methodological orientation toward establishing generalizability that has been deleterious in its effects on the scientific accumulation of knowledge in education.

Similarly, Carver (1978, p. 378) argued that

> Statistical significance testing has involved more fantasy than fact. The emphasis on statistical significance in educational research represents a corrupt form of the scientific method. Educational research would be better off if it stopped testing its results for statistical significance.

Lakatos (1978, p. 88) suggests that

> One wonders whether the function of statistical techniques in the social sciences is not primarily to provide a machinery for producing phoney corroborations and thereby a semblance of "scientific progress" where, in fact, there is nothing but an increase in pseudo-intellectual garbage.

Meehl (1978, p. 817; 823) is even more emphatic:

> I suggest to you that Sir Ronald [Fisher] has befuddled us, mesmerized us, and led us down the primrose path. I believe that the almost universal reliance on merely refuting the null hypothesis as the standard method for corroborating substantive theories in the soft [i.e., social science] areas is a terrible mistake, is basically unsound, poor scientific strategy, and one of the worst things that ever happended in the history of psychology.... I am not making some nit-picking statistician's correction. I am saying that the whole business is so radically defective as to be scientifically almost pointless.

The present paper evaluates the logic underlying various criticisms of statistical significance testing and makes specific recommendations for scientific and editorial practice that might better increase the knowledge base. The paper also evaluates the effects of contemporary significance testing practice on the literature. A final purpose of the paper is to explore why unconscious preferences for practices have emerged and why such practices are so impervious to change. Put differently, the purpose of the paper is to facilitate escape from some of the methodological paradigms that tend to govern thinking regarding the processes of scientific inquiry.

The Overarching Influence of Paradigms

Gage (1963, p. 95) defines paradigms as "models, patterns,

or schemata. Paradigms are not theories; they are rather ways of thinking or patterns for research." Tuthill and Ashton (1983, p. 7) suggest that

> A scientific paradigm can be thought of as a socially shared cognitive schema. Just as our cognitive schema provides us, as individuals, with a way of making sense of the world around us, a scientific paradigm provides a group of scientists with a way of collectively making sense of their scientific world.

Researchers tend to not be conscious of the influence of their paradigms on their research practices. As Lincoln and Guba (1985, pp. 19-20) note:

> If it is difficult for a fish to understand water because it has spent all of its life in it, so it is difficult for scientists... to understand what their basic axioms or assumptions might be and what impact those axioms and assumptions have upon everyday thinking and lifestyle.

Yet paradigms exert enormous influence, because they tend to tell researchers what they need to think about, and even more importantly, because paradigms also tell researchers the issues about which they need not think. As Patton (1975, p. 9) argues,

> Paradigms are normative; they tell the practitioner what to do without the necessity of long existential or epistemological consideration. But it this aspect of a paradigm that constitutes both

> its strength and its weakness--its strength in that it makes action possible, its weakness in that the very reason for action is hidden in the unquestioned assumptions of the paradigm.

Two examples of paradigm influence on thought can readily be cited. For example, Thompson (1986c, pp. 5-6) notes that researchers do not generally question their interpretation of "error" variance, although several interpretations are available. The second example is somewhat more troubling. Analysis of variance methods and their analogs (hereafter labelled OVA methods) are the most commonly employed analytic techniques in the social sciences (Goodwin & Goodwin, 1985) despite well known arguments against this preference (Cohen, 1968; Thompson, 1986a). Kerlinger (1986, p. 203) notes that "The analysis of variance is not just a statistical method. It is an approach and a way of thinking." The influence of an analysis of variance paradigm is part of the etiology that has led to overuse of "OVA" methods (Thompson, 1981).

Bakan (1966, p. 436) has suggested that, "When we reach a point where our statistical procedures are substitutes instead of aids to thought, and we are led to absurdities, then we must return to the common sense basis." Appreciating some of the arbitrary aspects of statistical significance testing may force recognition of the "significance" paradigm's potency; some knowledge of the historical origins of typical contemporary practice may help foster this insight.

Origins of Alpha Level Preferences

Olejnik (1984, p. 41) notes that researchers' preferences for the 0.05 alpha level have virtually assumed their own life force:

> Most hypotheses in the social sciences are tested at a 0.05 level of significance. While this criterion of significance is arbitrary, it has gained wide acceptance to the point where any hypothesis tested at a higher probability of a Type I error is viewed with considerable reservation.

Indeed, Fisher's original preference for tests at the 0.05 level was seemingly fairly casual, although it was apparently not random. Fisher's writings seem to reflect this:

> It is convenient to take this point [.05] as a limit in judging whether a deviation is to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant. (Fisher, 1925, p. 47)

> If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 per cent point), or one in a hundred (the one per cent point). Personally, the writer prefers to set a low standard of significance at the 5 per cent point, and ignore entirely all results which fail to reach this level. A scientific fact should be regarded as experimentally established only if a properly designed experiment <u>rarely fails</u> to give this level of significance. (Fisher, 1926, p. 504)

As Cochran (1976, p. 17) notes in a chapter in On the history of statistics and probability, "Fisher sounds fairly casual about the choice of 5% for the significance level, as the words 'convenient' and 'prefers' have indicated."

In some respects the preference for alpha being some small value is not itself arbitrary. As Lindquist (1953, pp. 68-70) has noted, the dangers to science of Type I error can be greater than Type II error, since the effect of a Type I error may be that "time and effort will be wasted on further experiments designed to determine the nature of the relationship." If a Type II error is made,

> ...the likely consequence is simply that we fail to follow up a true lead. In a sense this is not as serious as to have wasted time following up a false lead, since in the meantime we may be trying out other possible leads, all of which eventually had to be tried out anyway. (p. 68)

However, it is ironic that Fisher's preference for the level of alpha apparently was also influenced by paradigms. For example, Cowles and Davis (1982, p. 556) argue that Fisher was influenced by related scientific precedent particularly in astromony and that "Fisher then cannot be credited with establishing the value of the significance level."

Cowles and Davis (1982) also suggest that Fisher (and subsequent researchers) may have been influenced by a psychological dynamic involving "subjective probability" (Alberoni, 1962a, 1962b). The authors define "subjective probability" by noting that

> If, however, at some point the events begin to contradict the expectations they [individuals] have formed, they introduce cause and abandon the idea of chance. The point at which this rejection occurs depends largely on the degree of discrepancy and how it is interpreted by each individual. (Cowles & Davis, 1982, p. 557)

Cowles and Davis (1982) suggest that in their daily lives most people utilize deviations from chance that are only 5% probable as the limen for revising their psychological expectations regarding environmental events.

Criteria for Paradigm Evaluation

Since the purpose of a scientific methodological paradigm is to facilitate the expansion of the knowledge base, the significance testing paradigm must be evaluated against its ability to facilitate this end. However, the evaluation must avoid what has been termed the "is/ought" error (Hudson, 1969). As Strike (1979, p. 13) explains:

> To deduce a proposition with an "ought" in it from premises containing only "is" assertions is to get something in the conclusion not contained in the premises, something impossible in a valid deductive argument.

For example, arguing that quantitative research is poorly done is not sufficient reason to reject the paradigm unless it is demonstrated that the problem is inherent in the paradigm. Eisner (1983, p. 14) notes that

> The median experimental treatment time for seven of the 15 experimental studies that reported experimental treatment time in Volume 18 of the AERJ is 1 hour and 15 minutes. I suppose that we should take some comfort in the fact that this represents a 66 percent increase over a 3-year period. In 1978 the median experimental treatment time per subject was 45 minutes.

Eisner would have made an "is/ought" error if he had argued that, because quantitative researchers often are not doing a good job, that therefore we ought to abandon the quantitative research paradigm.

With respect to statistical significance testing, one common criticism derives as the conclusion from the two premises of a syllogism. First, it is suggested that test statistics often presume random sampling from the population and random assignment to treatment conditions. Second, it is correctly noted both that "it is rarely the case that investigators truly sample from a total population" (Shulman, 1981) and that random assignment to meaningful experimental groups is also quite rare in some areas of inquiry (Welch & Walberg, 1974, p. 113). The syllogism's conclusion is that therefore researchers ought to question the legitimacy of many applications of significance testing (Shaver & Norton, 1980).

One difficulty with this logic is that the first premise is not entirely true. Significance testing imposes a restriction that samples must be representative of a population, but does not

mandate that this end must be realized through random sampling. If researchers are unable to sample randomly, it may be possible to build a Cornfield-Tukey (1956) "bridge" from the sample to the population by comparing known sample characteristics with known population characteristics to build some warrant for an assumption of representativeness. Similarly, experimental designs presume group equivalence at the initiation of the inquiry, and not a particular method for realizing this end. It is noteworthy that Carlberg and Kavale (1980, p. 303) found in their meta-analytic study that the presence or absence of random assignment made virtually no difference in treatment effect sizes. McGinnis (1958, p. 413) summarizes the matter by noting that:

> No test of significance requires of itself that all correlated biases be removed, that is, that randomization be effected. It is true that this process assures that some requisite conditions of certain test statistics will be reasonably well met. Failure to randomize, however, in no way assures that such conditions will be violated. In general, then, the claim that all statistical tests of all hypotheses require the experimental procedure of randomization is unwarranted.

Misinterpretations of Significance Tests

Various misinterpretations of tests of statistical significance have been catalogued in books and articles (e.g., Carver, 1978). The misinterpretations are not inherent in the paradigm, except to the extent that paradigms by encouraging

practice without thought can encourage thoughtless practice.

A serious misinterpretation of test results occurs when researchers use statistical significance calculations to try to evaluate whether results are important. Significance tests do not require as input into calculations declarations of the researcher's value system, and therefore cannot contain as output information about the importance of results. Again, deductions may not contain in conclusions information not utilized in premises. As Daniel (1977, p. 425) explains,

> Whether or not the magnitude of the difference between Mu of A and Mu of B is of any practical importance is a question that cannot be answered by the statistical test. This is a question that only the researcher can answer after consideration of nonstatistical information.

Although most researchers are aware that this misinterpretation is problematic, paradigm influence can lull some researchers into not realizing just when they are making this error.

Another serious misinterpretation occurs when researchers conclude based on a significant sample result that the null hypothesis is necessarily false in the population. As Kish (1959, p. 336) observes:

> After finding a result improbable under the null hypothesis the researcher must not accept blindly the hypothesis of "significance" due to a presumed cause. Among the several alternative hypotheses is that of having discovered an improbable random event through sheer diligence.

Significance tests evaluate the calculated probability that a sample result originated in a population in which the null hypothesis is true. The tests do not establish to a certainty whether the sample results came from such a population, and the results do not establish a certainty as to whether the null is true or false in the population. Thus, one possible explanation for any "significant" result must always be that an unusual sample was selected from the population of possible samples, i.e., that sampling error produced the result.

Causes and Consequences of Bias Against Non-Significance

There can be little question but that the published literature historically reflects a major bias against statistically non-significant results. Empirical studies of journals confirm that few non-significant results are reported. For example, Sterling (1959) examined volumes from four psychology journals and found that only 1.9% of the articles reported non-rejection of primary null hypotheses. Greenwald (1975a, p. 12) reported that only 12.1% of the articles which he examined from the 1972 volume of the Journal of Personality and Social Psychology reported non-rejection of the primary null hypothesis.

Empirical studies of reviewer, author, and reader perceptions also corroborate the existence of a prejudice against non-significant results. For example, Cohen (1979) found that revised versions of published articles were rated more highly by counseling practitioners if the revisions reported statistically significant findings than if they reported statistically

nonsignificant findings. Similarly, Atkinson, Furlong and Wampold (1982, p. 189) reported that:

> In order to test for a statistical significance effect, 101 consulting editors of the Journal of Counseling Psychology and the Journal of Consulting and Clinical Psychology were asked to evaluate three versions, differing only with regard to level of statistical significance, of a research manuscript. The statistically nonsignificant and approach significance versions were more than three times as likely to be recommended for rejection than was the statistically significant version.

Most disturbing of all, Greenwald's (1975a, p. 5) study of 48 authors and 47 reviewers for the Journal of Personality and Social Psychology indicates that non-significant results discourage not only even submitting a manuscript, but continuing a line of inquiry as well.

> A major bais is indicated in the 0.49 (+/-.06) probability of submitting a rejection of the null hypothesis for publication (Question 4a) compared to the low probability of .06 (+/-.03) for submitting a nonrejection of the null hypothesis for publication (Question 5a). A secondary bias is apparent in the probability of continuing with a problem.

The bias against statistically non-significant results has traditionally even been a component of formal editorial policy.

For example, Melton (1962, p. 554), after 12 years as editor of the Journal of Experimental Psychology, noted that

> In editing the Journal there has been a strong reluctance to accept and publish results related to the principal concern of the researcher when those results were significant at the .05 level... It reflects a belief that it is the responsibility of the investigator in a science to reveal his effect in such a way that no reasonable man would be in a position to discredit the results by saying that they were the product of the way the ball bounces.

Similarly, the previous edition of the APA Publication Manual admonished that:

> Negative results lacking a theoretical context are basically uninterpretable. Even when the theoretical basis for the prediction is clear and defensible, the burden of methodological precision falls heavily on the investigator who reports negative results... Failure to replicate results of a previous investigator, using the same method but a different sample, is generally of questionable value. A single failure may merely testify to sampling error or to the conclusion that one one of the two samples had unique characteristics responsible for the reported effect, or the lack of effect. (American Psychological Association, 1974, 21)

As Atkinson, Furlong and Wampold (1982, p. 190) observe,

> If a single failure to replicate the findings of an earlier study can be written off as due to sampling error or the idiosyncracies of the sample population, what assurance is there that the results reported in the initial study were not a function of these same factors?

As Greenwald (1975b, p. 182) has argued, "very little is expected to be published on problems for which the null hypothesis is, to a reasonable approximation, true."

However, the null hypothesis of no difference is almost never exactly true in the population. As Savage (1957, pp. 332-333) notes, "Null hypotheses of no difference are usually known to be false before the data are collected." Similarly, Meehl (1978, p. 822) noted that, "As I believe is generally recognized by statisticians today and by thoughtful social scientists, the null hypothesis, taken literally, is always false." Hays (1981, p. 293) notes that, "There is surely nothing on earth that is completely independent of anything else. The strength of association may approach zero, but it should seldom or never be exactly zero."

Consequently, Greenwald (1975b, p. 190) notes that the significance testing paradigm (or at least many researchers' contemporary vision of the paradigm) diverts researchers away from important areas of scientific inquiry.

> It does appear that scientists' reputations are more readily established by looking for and finding new relationships that require new explanations

than by looking for and finding nonrelationships that would discredit old (particularly their own) explanations. But it is distressing that we accumulate new relationships and explanations without getting rid of corresponding numbers of old ones, that the new explanations are often difficult to make consistent with one another, and that we often fail to face important empirical and theoretical problems because our significance tests divert us from them.

Factors Affecting Statistical Significance

It is well known that sample size is a primary determinent of the statistical significance of results. As Kish (1959, p. 336, emphasis added) explains, "The results of statistical 'tests of significance' are functions not only of the magnitude of the relationships studied but also of the numbers of sampling units used (and the efficiency of the design)." Hays (1981, p. 293) argues that "Virtually any study can be made to show significant results if one uses enough subjects."

Thus, Nunnally (1960, p. 643) was not surprised that correlation coefficients based on data from 700 subjects all tended to be statistically significant: "If the null hypothesis is not rejected, it is usually because the N is too small. If enough data are gathered, the hypothesis will generally be rejected." Kaiser (1976) was not surprised when many substantively trivial factors were found to be statistically significant when data were available from 40,000 subjects. Bakan

(1966, p. 425) reports that, "The author had occasion to run a number of tests of significance on a battery of tests collected on about 60,000 subjects from all over the United States. Every test came out significant."

The fact that sample size affects significance tests is probably emphasized in every methods textbook on the market today. Yet, although most researchers know and doubtless believe the truism, researchers can still fail to appropriately apply this realization, again perhaps due to a paradigm's encouragement to not think. For example, Rosenthal and Gaito (1963) asked 9 doctoral faculty and 10 graduate students to indicate their confidence in results from studies. The subjects were more confident in larger sample results for a given alpha level. This is contrary to theory of the test logic since a given p calculated for a smaller sample should be more convincing. As Bakan (1966, p. 430) notes, "Indeed, rejecting the null hypothesis with a small n is indicative of a strong deviation from null in the population, the mathematics of the test of significance having already taken into account the smallness of the sample." Thompson (1986c, p. 14) concurs, noting that:

> Significance testing does inform scientific practice when a significant effect is detected given a small sample. The procedure also informs practice when sample size is large (i.e., there is good power against Type II and Type IV (Marascuilo & Levin, 1970) error) and a "significant" result is not realized. These are less likely occurences, and the researcher can argue, a fortiori, that the

results are especially noteworthy.

If researchers forced themselves to consciously evaluate how all the factors that affect statistical significance affect a result "in hand," then interpretation might surface above paradigm-facilitated avoidance of thought. Schneider and Darcy (1984, p. 575) present and discuss the various factors that affect statistical significance:

> The outcome of significance tests, however, is determined by at least seven factors, and actual impact is only one of these. These elements are: (1) Actual strength of impact; (2) Number of cases used in the study; (3) Variation among cases on relevant variables; (4) The complexity of the analysis (degrees of freedom); (5) The appropriateness of the statistical measures and tests used; (6) The hypotheses tested; (7) The significance level choosen.

Another (though related) factor that must be considered is the measurement reliability of the data. As Thompson (1986a, p. 919) explains,

> Statistically significant effects are theoretically possible only when variables are reliably measured. Reliability, in turn, is basically a function of variance. Longer tests tend to be more reliable than shorter tests only because tests with more items allow subjects to achieve scores that are more "spread out."

To be precise, only specific data collected from specific people on a specific occasion (not tests) have attributes involving reliability. Researchers who surmount the paradigm influence of classical test theory by relying on generalizability theory see this fairly readily (Brennan, 1983). As Sax (1980, p. 261) notes,

> It is more accurate to talk about the reliability of measurements (data, scores, and observations) than the reliability of tests (questions, items, and other tasks). Tests cannot be stable or unstable, but observations can. Any reference to the "reliability of a test" should always be interpreted to mean the "reliability of measurements or observations [i.e., a particular set of data] derived from a test."

Rowley (1976, p. 53) concurs, noting that "It needs to be established that an instrument itself is neither reliable nor unreliable." This suggests that the researcher bears a burden of proof for establishing that data collected from every occasion of use are reliable, since the fact that a measure is a constant across studies does not mean that the data from the measure will necessarily always be reliable. Meaningful interpretation of results requires considered evaluation of all the factors that may have yielded a given result.

The Importance of Effect Size Estimation

Textbook authors concur that sample size is a primary determinent of significance, especially given that the null hypothesis is almost always false in the population:

> It's our opinion that finding a "significant effect" really provides very little information, because it's almost certainly true that some relationship (however small) exists between any two variables. And in general finding a significant effect simply means that enough observations have been collected in the experiment to make the statistical test of the experiment powerful enough to detect whatever effect there is. (Loftus & Loftus, 1982, p. 498-499)

Alternatively, some textbooks even go so far as to suggest that researchers may sometimes want to use smaller sample sizes to avoid too much likelihood of a statistically significant result. For example, Hays (1981, p. 294) recommends asking, "Is the sample size large enough to give confidence that the big associations will indeed show up, while being small enough so that trivial associations will be excluded from significance?"

But most researchers are instinctively guided by an external validity paradigm that encourages the use of "as many subjects as possible." This is done "to 'wash out' the effects of individual differences or outliers, or to invoke the central limit theorem to meet distribution assumptions of parametric tests (B. Thompson, 1984, p. 18), or to maximize generalization" (Thompson, 1986c, p. 12). As Signorelli (1974, pp. 774-775) notes:

> The probability of obtaining significant results increases with sample size, regardless of the validity of the hypothesis under study. Yet, increasing sample size so as to obtain a "truly

representative" sample is a procedure recommended by the majority of statistical textbook authors.

Berkson's (1938, p. 527) conclusion some 50 years ago seems to often apply today, at least in areas of inquiry that more frequently involve larger sample sizes:

> I suppose it would be agreed by statisticians that a large sample is always better than a small sample. If, then, we know in advance the P that will result from an application of a chi-square test to a large sample, there would seem to be no use in doing it on a smaller one. But since the result of the former test is known, it is no test at all.

Unfortunately, the power of the significance testing paradigm to encourage researchers to not think influences many researchers to not go beyond the statistical significance of their results in interpretation. Yates (1951, pp. 32-33) noted this potential some time ago:

> [The use of tests of significance] has caused scientific research workers to pay undue attention to the results of the tests of significance they perform on their data, particularly data derived from experiments, and too little to the estimates of the magnitude of the effects they are investigating... The emphasis on tests of significance, and the consideration of the results of each experiment in isolation, have had the

> unfortunate consequence that scientific workers have often regarded the execution of a test of significance on an experiment as the ultimate objective. Results are significant or not significant and this is the end of it.

Shaver (1980, p. 13) has also noted the influence of the significance testing paradigm:

> Making decisions about the size an effect must reach to be important is usually not easy in educational research, for a number of reasons. One is that current research paradigms have not conditioned us to think that way.

A variety of effect size statistics have been proposed. They range in their sophistication from eta squared (Hays, 1981, p. 349), sometimes called the correlation ratio and readily available in the univariate case, to a corrected version of the multivariate omega squared statistic (Tatsuoka, 1973, p 30). These statistics indicate how much variance in the dependent variable(s) is accounted for by the treatment conditions or the predictor variables.

The statistics are important. As Loftus and Loftus (1982, p. 499) suggest, "it is our judgment that accounting for variance is really much more meaningful than testing for significance." Craig, Eison and Metze (1976, p. 282) argue that "Since psychology's task is not one of identifying 'significant' differences but 'meaningful' relationships among variables, it would appear that dealing only with significant differences is not enough." Hays (1981, p. 293) explains the consequences of

interpreting statistically significant results derived in large samples when effect size is not calculated and actually is small:

> This kind of problem occurs when people pay too much attention to the significance test and too little to the degree of statistical association the findings represent. This clutters up the literature with findings that are often not worth pursuing.

When researchers employ correlational analyses, the squared correlation coefficient is an effect size estimate and is available as part of the analysis. Unfortunately, empirical studies indicate that researchers do not usually report effect size estimates when non-correlational analyses are conducted. For example, McNamara (1978, p. 48) reported that

> A reanalysis of research studies reported in the first twelve volumes of the Educational Administration Quarterly led to the identification of 31 inquiries that employed a univariate analysis of variance as the basic statistical model for testing experimental hypotheses. In almost all cases these inquiries systematically avoided any reference to either practical significance or [effect size] tests for the significance of relationships.

This would be troubling only as a theoretical difficulty except that in practice many published studies reporting and interpreting significant results actually involve smaller effects sizes. Craig, Eison and Metze (1976, p. 281) empirically

evaluated effect sizes from 50 articles from three major psychology journals and found that:

> There is a great deal of variation among reported significant statistical outcomes in terms of the amount of shared variation present between the independent and dependent variables... General studies which employed large samples tended to have a small amount of shared variance with the opposite being observed with smaller samples.

To illustrate the problems originating from this situation, Wampold, Furlong and Atkinson (1983, p. 462) cite one study in which the researchers "attributed substantive importance to a statistically significant result that had an effect size of 0.05, whereas" in another study researchers "explained away a statistically nonsignificant result that had an effect size of 0.17."

The failure to utilize effect size estimates in interpretation may partially be due to the influence of the significance testing paradaigm which influences attention away from effect size. The failure to consult effect sizes also may be due to another paradigm problem involving researchers not recognizing that all analyses actually test association. As McNamara (1978, p. 50) notes, "Rejection of the hypothesis of no difference between population means is tantamount to the assertion that the independent variables do have some statistical association with the criterion scores."

This realization should be inescapable for researchers who overcome paradigm restrictions and think about the implications

of their awareness that all univariate tests can be conducted using multiple correlation analysis (Cohen, 1968) and that all univariate and all multivariate parametric tests are special cases of canonical correlation analysis (Knapp, 1978; Thompson, 1984, p. 1). Thompson (1985) provides concrete heuristic examples in which a small data set is analyzed using a variety of parametric analyses in order to demonstrate these identities.

Regrettably, many researchers still prefer analysis of variance methods that do not routinely provide effect size estimates. Thompson (1986c, p. 17) explains a possible origin for this preference:

> Many quantitative researchers prefer experimental designs because the designs allow somewhat more warranted confidence in the internal validity of conclusions about causality. The fact that OVA methods are appropriate when predictor variables such as experimental assignment naturally occur at the nominal level of scale has stimulated some researchers to unconsciously associate the consequences of design selection with OVA methods.

Researchers must see that the calculation of an effect size as part of analysis does not make the design correlational any more than the use of an OVA technique as part of analysis makes the design experimental.

The Importance of Power Estimation

It is well known that possible explanations for "non-significant" results include the possibility that the researcher

employed too few subjects to detect existing effects. The probability of this explanation can be evaluated by conducting a power analysis to estimate Type II error probability (Cohen, 1977). Fagley and McKinney (1983, p. 298) note that "Studies reporting nonsignificant findings contribute to the body of knowledge in a field only if their power is high." Fagley (1985, p. 391) notes that "Nonsignificant results can be a potential contribution to knowledge only when the power of the statistical tests was high and are ambiguous at best when the power of the statistical tests was low."

Researchers who do not employ power analysis in order to determine required sample sizes may fail to obtain significance as an artifact of inadequate sample size. Furthermore, researchers who publish non-significant results without reporting their power calculations force the reader to conduct the analysis in order to escape the quandry of whether results are apparently genuine or merely reflect Type II error. However, relatively few articles report these analyses (Olejnik, 1984). This might not be troubling if most researchers chanced into having adequate power. However, Woolley (1983, p. 710) found that "Fully 91 percent of the 100 [medical] articles analyzed had less than a 50-50 chance of detecting a 'small' treatment effect."

This result is particularly noteworthy given that most studies tend to find what Cohen (1977, pp. 79-80) has characterized as a medium effect size. For example, Olejnik (1984, p. 43) reviewed a series of meta-analytic studies and found that the median effect size in research tends to be about

0.3. Glass (1979) concurs, noting that:

> In none of the dozen or so research literatures that we have integrated in the past five years have we ever encountered a cross-validated multiple correlation between study findings and study characteristics that was larger than approximately 0.60. That is, I haven't seen a body of literature in which we can account for much more than a third of the variability in the results of studies.

So many articles almost all reporting statistically significant results with varying effect sizes raises the spectre of a literature replete with Type I errors, notwithstanding small alpha levels. It is as if researchers pick small alpha levels to avoid Type I errors, but that the bias against non-significant results and the reluctance to think about effect size and power then creates a dynamic at a higher level that biases toward whatever Type I errors do occur. Even though researchers recognize the seriousness of Type I error as regards scientific progress (Lindquist, 1953, pp. 68-70), the higher-level bias may make a strong correction lessening this protection. The problem is serious because a single study that reports a signficant result based on actual (albiet unlikely) Type I error can begin a cascade of replication studies that then tend to be submitted and published only if they are supportive of initial results. Greenwald (1975a, pp. 13-15) cites several such "epidemics" from the literature.

Although the literature may be biased in favor of Type I error, at least some comfort might be taken in knowledge that the

bias does lessen reporting of Type II errors. By definition, a Type II error cannot occur in a study in which the null hypothesis is not rejected. However, in a literature without a bias toward significance and Type I error, power analyses would be an important component of quantitative scientific inquiry.

The Importance of Replication in Science

Researchers have increasingly recognized the critical nature of replication as the ultimate test of scientific findings and some have argued that replicability should replace significance testing as part of a new logic of truth testing. As Gold (1969, p. 43) notes:

> Random sampling is by no means a necessary criterion for establishing the validity of a proposition statistically expressed. The validity we seek in social science research can come only from repeated observation under varying conditions of population.

Similarly, Neale and Liebert (1986, p. 290) argue that

> No one study, however shrewdly designed and carefully executed, can provide convincing support for a causal hypothesis or theoretical statement in the social sciences... How, then, does social science theory advance through research? The answer is, by collecting a diverse body of evidence about any major theoretical proposition.

Some researchers have suggested that replication is particularly important given contemporary practice in the social

sciences:

> Since effect sizes in the social sciences tend to be small and sample sizes often cannot be increased greatly, a reasonable alternative for maintaining statistical power is to accept an increased chance of a Type I error. Over replications of the study, true effects would be separated from Type I errors. (Olejnik, 1984, p. 47)

However, the significance testing paradigm may divert attention away from the realization that replication is vitally important in science. As Schwartz and Dalgleish (1982, pp. 290-291) note

> Although the notion that science proceeds gradually, accumulating evidence from a sequence of corroborative studies, is accepted by all psychologists, the classical (that is, Fisherian) statistics they routinely use to analyze their results are based on a rather different inferential model. The inference model underlying classical statistics assumes that only a single experiment out of a hypothetical population of experiments is actually conducted... Replications play no role in this reasoning except when they are expressly included in the statistical model (as in repeated-measures analysis of variance, for example).

One implication of these views is that researchers should more often consider employing "hold out" samples to cross-validate the results from their analyses. The methods for such

analyses are well known for correlational analyses, including multiple correlation analysis (Huck, Cormier & Bounds, 1974, p. 159), factor analysis (Gorsuch, 1983, p. 334; Thompson, 1986b), and canonical correlation analysis (Thompson, 1984). The same types of methods might be employed by a researcher using other analyses as part of a mini-replication study. However, these cross-validation results must be interpreted with some caution because, as Thompson (1984, p. 46) notes, "all procedures are 'liberal' estimates of invariance when one data set is split into subgroups, because the two subgroups came from one sample and the subgroups and their parameter estimates are therefore interdependent."

Recommendations for Improved Editorial Policy and Practice

The previous discussion naturally culminates in recommendations for improved editorial policy and for improved research practice. Several recommendations seem warranted.

First, researchers should overcome hesitancy to submit "nonsignificant" results for publication, and editors and reviewers ought to abandon a prejudice against such results. However, when nonsignificant results are reported and Type II error is thus a possibility, researchers ought to report power analyses indicating that reasonable confidence can be vested in an unambiguous result. Some researchers find it difficult to conduct power analyses in order to determine required sample sizes, since the anticipated effect size must be declared in advance of data collection. The expected outcome can be determined by consulting previous studies, or the expected effect

size can be identified by specifying the minumum effect that the researcher feels must be realized in order to achieve a practical result, i.e., a result that is persuasive regarding policy decisions or the validity of theory. However, even in new areas of inquiry, the consistency of effect sizes reported in various literatures (Glass, 1979; Olejnik, 1984) suggests that, failing all else, Cohen's (1977) medium effect size designations might be employed for the purposes of power analyses.

Second, effect sizes ought to be routinely reported, regardless of whether the design is experimental or correlational and of whether the analysis is correlational or involves OVA methods. Researchers should pay serious attention to the substantive importance of their results as against abandoning these decisions to statistics in some unconscious but headlong escape from freedom. Many researchers may claim awareness that statistical significance cannot be interpreted as importance, but the actions of researchers who compute test statistics and not effect sizes belie the claimed understanding, because these researchers are relegating value judgments to their probability calculations.

The following statement is a model for relevant sections of editorial board policy that might result in an improved literature:

1. Given the influence of sample size on tests of statistical significance, results are less ambiguous when effect sizes are reported. To facilitate the reader's evaluation of results, authors should compute and report effect sizes for their various tests. Furthermore, the author's

interpretation of results should seriously consider effect size as a critical aspect of results. Interpretations that consider comparisons with effect sizes in previous research in related areas of inquiry are particularly encouraged.

2. The journal encourages the submission of manuscripts reporting statistically non-significant results when results are unambiguous. Since Type II error is a possibility in such cases, authors should present their estimates of the study's power against Type II error, just as studies reporting significant results report protection levels against Type I error.

Research practice might also improve if researchers overcame paradigm influences and realized that the descriptive and the inferential elements of their analyses are distinct. Even if significance test statistics were entirely abandoned, this would not mean that statistical procedures such as analysis of variance or the correlation coefficient would necessarily have to be abandoned. For example, a researcher might be most interested in addressing the question, What proportion of variance does main effect A have on the dependent variable(s)? In an analysis of variance, a proportion might be calculated by dividing the main effect sum of squares by the total sum of squares. In a multivariate analysis of variance, lambda might be subtracted from one. Thus, statistical analyses can be implemented and effect sizes can be calculated even if the associated test statistics are never computed. It is also helpful to distinguish

the statistical assumptions that underlie a descriptive statistical analysis from the assumptions underlying the test statistics that can be used in the analysis; when the test statistics will not be applied the researcher need only meet the assumptions for the descriptive elements of the analysis (cf. Thompson, 1984, pp. 16-18). Critics of statistical methodology who reject significance testing in all its forms are attacking only one element of the corpus--an element that theoretically could be amputated without injury to the remaining elements. Thus, critics of the quantitative research paradigm who build their case entirely on criticisms of significance testing do not build a compelling case for their position.

On a more general level, the previous discussion suggests a recommendation that social scientists should continually work to overcome the strictures of their various paradigms. This will require recognition that significance testing informs interpretation only in special cases, e.g., significant results achieved with small samples. This may lead to better interpretation of results. And it will require recognition that all inquiry involves questions about association even when OVA analyses are employed. This may lead to more considered selection of analytic methods. It will require recognition that reliability is a characteristic inuring to data and not tests. This may lead to better measurement practices. And it will require a recognition that replication is the ultimate test of the generalizability of findings. These and related realizations may lead researchers toward the development of better theory. As Dar (1987, p. 149) notes, "When passing null hypothesis tests becomes

the criterion for successful predictions, as well as for journal publications, there is no pressure on the psychology researcher to build a solid, accurate theory; all he or she is required to do, it seems, is produce 'statistically significant' results."

References

Alberoni, F. (1962). Contribution to the study of subjective probability. Part I. Journal of General Psychology, 66, 241-264. (a)

Alberoni, F. (1962). Contribution to the study of subjective probability: Prediction. Part II. Journal of General Psychology, 66, 265-285. (b)

American Psychological Association. (1974). Publication manual of the American Psychological Association (2nd ed.). Washington, D.C.: Author.

Atkinson, D. R., Furlong, M. J., & Wampold, B. E. (1982). Statistical significance, reviewer evaluations, and the scientific process: Is there a (statistically) significant relationship? Journal of Counseling Psychology, 29, 189-194.

Bakan, D. (1966). The test of significance in psychological research. Psychological Bulletin, 66, 423-437.

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. Journal of the American Statistical Association, 33, 526-542.

Brennan, R. L. (1983). Elements of generalizability theory. Iowa City, IA: The American College Testing Program.

Carlberg, C. G., & Kavale, K. (1980). The efficacy of special versus regular class placement for exceptional children: A meta-analysis. Journal of Special Education, 14, 295-309.

Carver, R. P. (1978). The case against statistical significance testing. Harvard Educational Review, 48, 378-399.

Cochran, W. G. (1976). Early development of techniques in comparative experimentation. In D. B. Owen (Ed.), On the history of statistics

and probability. New York: Dekker, pp. 2-25.

Cohen, J. (1968). Multiple regression as a general data-analytic system. Psychological Bulletin, 70, 426-443.

Cohen, J. (1977). Statistical power analysis for the behavioral sciences (rev. ed.). New York: Academic Press.

Cohen, L. H. (1979). Clinical psychologists' judgments of the scientific merit and clinical relevance of psychotherapy outcome research. Journal of Consulting and Clinical Psychology, 47, 421-423.

Cornfield, J., & Tukey, J. W. (1956). Average values of mean squares in factorials. Annals of Mathematical Statistics, 27, 907-959.

Cowles, M., & Davis, C. (1982). On the origins of the .05 level of statistical significance. American Psychologist, 5, 553-558.

Craig, J. R., Eison, C. L., & Metze, L. P. (1976). Significance tests and their interpretation: An example utilizing published research and omega-squared. Bulletin of the Psychonomic Society, 7, 280-282.

Cronbach, L. J. (1975). Beyond the two disciplines of psychology. American Psychologist, 30, 116-127.

Daniel, W. W. (1977). Statistical significance versus practical significance. Science Education, 61, 423-427.

Dar, R. (1987). Another look at Meehl, Lakatos, and the scientific practices of psychologists. American Psychologist, 42, 145-151.

Eisner, E. W. (1983). Anastasia might still be alive, but the monarchy is dead. Educational Researcher, 12, 23-24.

Fagley, N. S. (1985). Applied statistical power analysis and the interpretation of nonsignificant results by research consumers.

statistics in research. New York: Harper & Row.

Hudson, W. D. (1969). The is/ought question. London: MacMillan.

Kaiser, H. F. (1976). [Review of Factor analysis as a statistical methos]. Educational and Psychological Measurement, 36, 586-589.

Kerlinger, F. N. (1986). Foundations of behavioral research (3rd ed.). New York: Holt, Rinehart & Winston.

Kish, L. (1959). Some statistical problems in research design. American Sociological Review, 24, 328-338.

Knapp, T. R. (1978). Canonical correlation analysis: A general parametric significance testing system. Psychological Bulletin, 85, 410-416.

Lakatos, I. (1978). Falsification and the methodology of scientific research programs. In J. Worrall & G. Currie (Eds.), The methodology of scientific research programs: Imre Lakatos philosophical papers (Vol. 1). Cambridge, Englad: Cambridge University Press, pp. 8-101.

Lincoln, Y. S., & Guba, E. G. (1985). Naturalistic inquiry. Beverly Hills: SAGE.

Lindquist, E. F. (1953). Design and analysis of experiments in psychology and education. Boston: Houghton Mifflin.

Loftus, G. R., & Loftus, E. F. (1982). Essence of statistics. Monterey, CA: Brooks/Cole.

McGinnis, R. (1958). Randomization and inference in sociological research. American Sociological Review, 23, 408-414.

McNamara, J. F. (1978). Practical significance and statistical models. Educational Administration Quarterly, 14, 48-63.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir

Karl, Sir Ronald, and the slow progress of soft psychology. Journal of Consulting and Clinical Psychology, 46, 806-834.

Melton, A. (1962). Editorial. Journal of Experimental Psychology, 64, 553-557.

Morrison, D. E., & Henkel, R. E. (1970). The significance test controversy--A reader. Chicago: Adeline.

Neale, J. M., & Liebert, R. M. (1986). Science and behavior: An introduction to methods of research (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Nunnally, J. (1960). The place of statistics in psychology. Educational and Psychological Measurement, 20, 641-650.

Olejnik, S. F. (1984). Planning educational research: Determining the necessary sample size. Journal of Experimental Education, 53, 40-48.

Patton, M. Q. (1975). Alternative evaluation research paradigm. Grand Forks, ND: University of North Dakota Press.

Rosenthal, R., & Gaito, J. (1963). The interpretation of levels of significance by psychological researchers. Journal of Psychology, 55, 33-38.

Rowley, G. L. (1976). The reliability of observational measures. American Educational Research Journal, 13, 51-59.

Savage, R. J. (1957). Nonparametric significance. Journal of the American Statistical Association, 52, 331-344.

Sax, G. (1980). Principles of educational and psychological measurement and evaluation (2nd ed.). Belmont, CA: Wadsworth.

Schneider, A. L., & Darcy, R. E. (1984). Policy implications of using significance tests in evaluation research. Evaluation Review, 8, 573-582.

Schwartz, S., & Dalgleish, L. (1982). Statistical inference in personality research. Journal of Research in Personality, 16, 290-302.

Selvin, H. C. (1957). A critique of tests of significance in survey research. American Sociological Review, 22, 519-527.

Shaver, J. P. (1979). The productivity of educational research and the applied-basic distinction. Educational Researcher, 8, 3-9.

Shaver, J. P. (1980). Readdressing the role of statistical tests of significance. Paper presented at the annual meeting of the American Educational Research Association, Boston, MA. (ERIC Document Reproduction Service No. ED 191 904)

Shaver, J. P., & Norton, R. S. (1980). Randomness and replication in ten years of the American Educational Research Journal. Educational Researcher, 9, 9-15.

Shulman, L. S. (1981). Disciplines of inquiry in education: An overview. Educational Researcher, 10, 5-12, 23.

Signorelli, A. (1974). Statistics: Tool or master of the psychologist? American Psychologist, 29, 774-777.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance--or vice versa. Journal of the American Statistical Association, 54, 30-34.

Strike, K. A. (1979). An epistomology of practical research. Educational Researcher, 8, 10-16.

Tatsuoka, M. M. (1973). An examination of the statistical properties of a multivariate measure of strength of relationship. Urbana: University of Illinois. (ERIC Document Reproduction Service No. ED 099 406)

Thompson, B. (November, 1981). The problem of OVAism. Paper presented at the annual meeting of the Mid-South Educational Research Association, Lexington, KY. [Order document #03980 from National Auxillary Publication Service, P.O. Box 3513, Grand Central Station, NY, NY 10017]

Thompson, B. (1984). Canonical correlation analysis: Uses and interpretation. Beverly Hills: SAGE.

Thompson, B. (1985). Heuristics for teaching multivariate general linear model concepts. Paper presented at the annual meeting of the American Educational Research Association, Chicago. (ERIC Document Reproduction Service ED 262 073)

Thompson, B. (1986). ANOVA versus regression analysis of ATI designs: An empirical investigation. Educational and Psychological Measurement, 46, 917-928. (a)

Thompson, B. (April, 1986). A partial test distribution for cosines among factors across samples. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA. (b)

Thompson, B. (April, 1986). The place of qualitative research in contemporary social science: The importance of post-paradigmatic thought. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA. (c)

Tuthill, D., & Ashton, P. (1983). Improving educational research through the development of educational paradigms. Educational Researcher, 12, 6-14.

Wampold, B. E., Furlong, M. J., & Atkinson, D. R. (1983). Statistical significance, power, and effect size: A response to the reexamination of reviewer bias. Journal of Counseling Psychology,

30, 459-463.

Welch, W. W., & Walberg, H. J. (1974). A course evaluation. In H. J. Walberg (Ed.), Evaluating educational performance: A sourcebook of methods, instruments, and examples. Berkeley: McCutchan, pp. 113-124.

Woolley, T. W. (1983). A comprehensive power-analytic investigation of research in medical education. Journal of Medical Education, 85, 710-715.

Yates, F. (1951). The influence of Statistical methods for research workers on the development of the science of statistics. Journal of the American Statistical Association, 46, 19-34.